

On The Stability of Video Detection and Tracking

Hong Zhang
Chinese University of Hong Kong
fykalviny@gmail.com

Naiyan Wang
TuSimple LLC
winsty@gmail.com

Abstract

In this paper, we study an important yet less explored aspect in video detection and multi-object tracking – stability. Surprisingly, there is no prior work that tried to quantify it. As a consequence, we start our work by proposing a novel evaluation metric for video detection which considers both stability and accuracy. For accuracy, we extend the existing accuracy metric mean Average Precision (mAP). For stability, we decompose it into three terms: fragment error, center position error, scale and ratio error. Each error represents one type of stability. Furthermore, we demonstrate that the stability metric has low correlation with accuracy metric. Thus, it indeed captures a different perspective of quality in object detection. Lastly, based on this metric, we evaluate several existing methods for video detection, and show how they affect accuracy and stability. We believe our work can provide guidance and solid baselines for future researches in related areas.

1. Introduction

Object detection refers to the problem that localizes and classifies the objects of interest in the image. It serves as a fundamental task for other high level tasks such as human computer interaction. In its early stage, most researches focus on certain object detection, such as face [44, 36, 32], hand [21, 30] or pedestrian [25, 31], etc. For general object detection, early method such as Deformable Part Model (DPM) [11] relies on hand-crafted features and deliberately designed classifiers. Nowadays, with the advancement of deep learning technique, current state-of-the-art paradigm shifts to data driven end-to-end learning. Some representative methods include region based methods [14, 13, 35, 16] and direct regression methods [34, 28].

Among all object detection methods, most of them focused on still images. Thanks to Convolutional Neural Networks (CNNs), great success has been achieved in this field. However, these methods still face difficulties such as motion blurs, occlusions and small scale objects. Recently, a series of works have been proposed by introducing cascaded net-

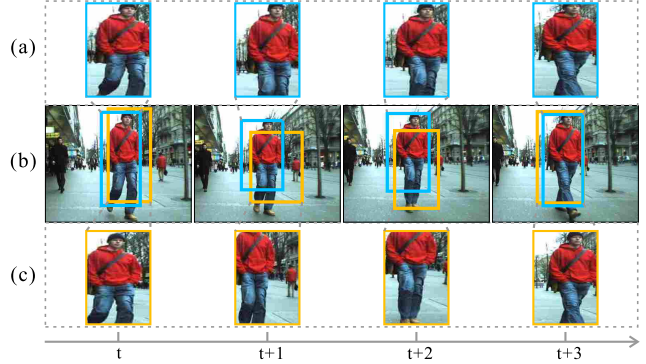


Figure 1. A pivot illustration to demonstrate the insufficiency of current VID evaluation. Here are results of two detectors in one video with the same mAP. (a) Stable trajectory. Though missing the feet of the pedestrian, the detector is consistent on the region it includes, and almost centered. (b) Consecutive frames with results of two detectors. (c) Unstable trajectory. Predicted bounding boxes contain different parts of object and jitter in a large region.

works [45], multi-region and multi-layer feature concatenation [22, 12] and context embedding [2] to alleviate the impact of these issues.

Another equally important yet less explored line is Video Detection (VID). Video detection could utilize the temporal context within consecutive frames to solve the aforementioned difficulties in a single frame. Though VID shares a lot of similarities with Multi-Object Tracking (MOT), the key difference lies in that VID does not output associations between frames. To push this field forward, ImageNet [8] introduced the video detection challenge recently. In just two years, we have witnessed that the performance improves from 67.82 to 80.83 in mean Average Precision (mAP) rapidly. Despite its rapid development, we have observed some disturbing facts regarding to the evaluation of VID algorithms. For example, in Fig. 1, two methods get the same results on this trajectory based on mAP. Nevertheless, (a) is obviously better than (c) for human judgment. We would like to ask: *What is the missing component in current VID evaluation?*

Our answer is stability. Currently, the evaluation of VID simply aggregates the results of still images. The evaluation

is purely based on the Intersection Over Union (IOU) metric which is limited to an individual image. The evaluation across different frames is not considered. Although other closely related tasks such as MOT evaluate the association accuracy, the stability along each trajectory is still ignored. Nevertheless, it usually plays an important role in practice. For example, in an Advanced Driving Assistance System (ADAS), we need to use the change rate of bounding box to estimate the Time To Collision (TTC) [7] and relative speed and distance [39], which is at the core of the safety warning system. If the center or scale of the bounding box jitters around the object of interest, it is obvious that the estimated TTC, speed and distance are inaccurate and unstable. Unfortunately, there is no evaluation metric to quantify such phenomenon accurately.

In this paper, we highlight the importance of stability in video detection in addition to its accuracy counterpart. In particular, the new metric evaluates the results from two aspects: The first one is detection accuracy. It extends the mAP in still image object detection by considering all IOU thresholds. This metric evaluates whether the objects of interest are precisely localized and classified. The second one is detection stability. We move the paradigm of evaluation from bounding box centric to trajectory centric. This metric assesses the stability along each trajectory, which includes temporal continuity, center position stability, scale and ratio stability, respectively. Guided by the new evaluation metric, we also benchmark several existing methods to improve VID task. The empirical results also reveal an interesting finding: These two metrics are less correlated. They capture different aspects of VID quality. We wish these methods could be served as effective baselines for future researches.

To summarize, our contributions are in the following three folds:

- We propose a novel evaluation metric to assess the performance of VID methods. The proposed metric considers a crucial yet usually ignored aspect of VID – stability. It is also an important supplement to current MOTA metric which is commonly used in MOT evaluation.
- We empirically demonstrate that the stability metric has low correlation with existing accuracy metric, thus it is meaningful to evaluate both of them.
- We evaluate some existing baselines under our new evaluation metric, and show trade-offs between them. We wish these benchmarks could lighten further research directions in the field.

2. Related Work

Object detection has a long history in computer vision community. Before deep learning age, the conventional

methods heavily relied on hand-crafted features and carefully designed pipelines. One representative work is the Deformable Part Model (DPM) [11]. Afterwards, instead of using sliding window, researchers proposed to first generate object proposals that may contain objects, and then classify them into different categories. Widely used methods to generate proposals include those based on super-pixels grouping, *e.g.* Selective Search [42], MCG [1] and those based on sliding window and edge features, *e.g.* EdgeBoxes [46].

Some early works tried to apply CNN to object detection include [37, 9, 40]. However, their performance does not significantly outperform the conventional methods. Region CNN (RCNN) [14] is a milestone in object detection. Briefly speaking, it adopts CNNs to extract features of region proposals and score them via learned classifiers. Recently, subsequent works such as Fast RCNN [13] and Faster RCNN [35] significantly accelerate it. The core idea behind these work is to reuse existing computation. For example, Fast RCNN uses ROI pooling to share feature maps in the feature extraction for each proposal; Faster RCNN shares the convolutional layers for proposal generation. By using the techniques above, the inference speed of one image decreases from 10s to 0.2s without sacrificing any performance. To further speed up, several real-time methods [34, 28] were proposed. These methods cast detection into direct regression problem. Namely, the network jointly predicts bounding boxes and confidence scores in an end-to-end manner.

Beyond object detection in still images, video detection aims to detect objects in video sequences. Until recently, a few researchers have done several pioneering works. The keys of these methods lie in utilizing temporal context information. In [18, 19], Kang *et al.* tried to use object class correlation and motion propagation to reduce false positives and false negatives first, and then train a temporal convolution network to rescore detections based on the tubelets generated by visual tracking. Han *et al.* [15] used sequence Non-Maximal Suppression (NMS) to rescore the detection results. Very recently, Tripathi *et al.* [41] trained a Recurrent Neural Network (RNN) on the initial detections results to refine them. Unfortunately, all these methods are limited in post-processing stage. Few works tried to integrate the temporal context in an end-to-end manner. Interestingly, for a closely related area – video segmentation, [10, 43] adopted Convolutional Long Short Term Memory (ConvLSTM) to capture both temporal and spatial information in one unified model. For the evaluation of video segmentation, a recent work [24] evaluated the temporal consistency as well as the accuracy in single frame. We believe video detection is still at its early stage compared with video segmentation. There are still a lot of issues that need to be addressed in the future. Among those, stability is the foremost one that needs to be investigated.

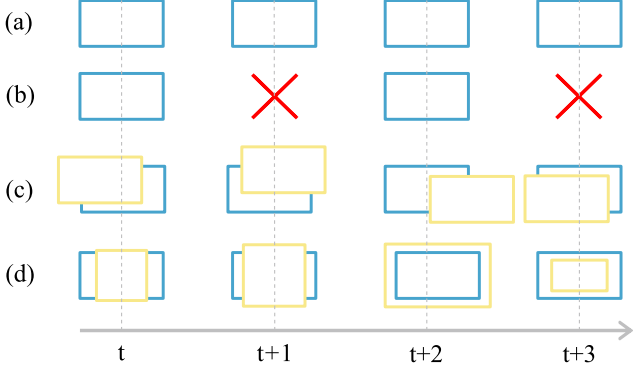


Figure 2. An illustration of four trajectories. (a) Ground-truth trajectory. (b) Interrupted trajectory. (c) Center position jittering trajectory. (d) Scale and ratio jittering trajectory.

Video detection is also a prerequisite for most MOT algorithms. They usually assume the detection results exist, and focus on associating and refining the detections. MOT algorithms can be roughly categorized into two types: (Nearly) online algorithms [5, 6, 20] try to associate existing targets with detections in recent frames and output results immediately after receiving the input image. Meanwhile, offline algorithms [3, 29] read in all frames, and then output the results afterwards. Offline algorithms could utilize the later frames to refine the results in early frames. To this end, the evaluation of MOT also emphasizes more on association in videos. Its commonly used metric is called Multi-Object Tracking Accuracy (MOTA) [4]. MOTA focuses on evaluating the association accuracy. In MOTA, besides false positive and false negative in detection, it will penalize the identity switch of detected objects. In contrary to MOT, video detection does not output associations between frames. Note that, though we do not require the algorithms to output associations, our proposed metric can also be applied to MOT. It captures a different aspect of trajectory quality in addition to existing MOTA metric.

3. Our Evaluation Metric

In this section, we present our novel evaluation metric for VID. The metric consists of two parts: The first part is detection accuracy which evaluates whether the objects are precisely localized and classified. The other part considers the detection stability for each trajectory both temporally and spatially. Temporal stability is to measure the integrity of a trajectory, while spatial stability is to measure how much the detection bounding box jitters around the ground-truth in a trajectory. An illustration for the stability metric is presented in Fig. 2. We will elaborate the details in the following sections.

3.1. Detection Accuracy

In each frame, the performance of a detector can be measured by the overlap rate between the ground-truths and the predicted bounding boxes. The overlap rate is defined as the area of the intersection of two bounding boxes over their union (IOU). Given an IOU threshold to be true positive, we can calculate the precision and recall curve. Average Precision (AP) is the Area Under Curve (AUC) of precision and recall curve. Then we vary the IOU threshold from 0 gradually to 1, and we calculate the AUC of AP curve as detection accuracy. A larger AUC value indicates better accuracy. If the dataset contains more than one class, we simply use the mean AUC (mAP) over all classes as the final detection accuracy.

3.2. Detection Stability

To evaluate the stability, we need to compare all the detections with respect to its corresponding ground-truth trajectory. In other words, we need the associations for the ground-truth, however the evaluated algorithm does not need to output them. We first apply Hungarian algorithm [23] to find the best matching between the output detections and ground-truths. IOUs between them are treated as the weights of the bipartite graph. The detections that are not paired to any ground-truth are excluded in stability evaluation. At last, we then average the stability errors of all categories as the final result. Eq. (1) is the formulation for detection stability:

$$\Phi = E_F + E_C + E_R, \quad (1)$$

where E_F is the fragment error, E_C is the center position error, E_R is the scale and ratio error. As in detection accuracy, we also change the IOU threshold of true positive to examine the trends between the errors and thresholds. For each IOU threshold, we can draw its corresponding error vs. recall curve. We use the AUC of this curve as the stability error at a certain threshold. At last, we treat the AUC of the IOU threshold vs. stability error curve as the final stability error. Lower value means higher stability in a video. These three components are presented in the following.

3.2.1 Fragment Error

In fragment error we want to evaluate the integrity of the detections in one trajectory. Particularly, the results of a stable detector should be consistent (always report as a target or always not). It should not frequently change its status throughout the trajectory. Formally, Let N be the total number of trajectories in a video sequence. t_k is the total length of the k^{th} trajectory, and f_k is the number of status change. One status change is defined in the scenario that the object is detected in previous frame but missed in current frame

and vice versa. Then the fragment error is defined as:

$$E_F = \frac{1}{N} \sum_{k=1}^N \frac{f_k}{t_k - 1}. \quad (2)$$

As a special case, we define fragment error of a trajectory with length one to be 0. The fragment error is minimized when the detector can always localized the object accurately along the ground-truth trajectory or never been, while it is maximized when the object is detected alternately. It is also noteworthy that there is also one metric with same name [26] in the evaluation of MOT, however there are two key differences: 1) We find the best matching between detections and ground-truths, while fragment error in MOT uses the association reported by the algorithms. 2) Our metric is normalized by the trajectory length, while the one in MOT does not. This makes the metric comparable across different videos with different numbers of trajectories.

3.2.2 Center Position Error

In center position error, we evaluate the stability of the center positions of the detections in one trajectory. This metric is illustrated in Fig. 2(c). A good detector should keep the centers of its outputs stable, instead of jittering around the centers of the ground-truths. We evaluate the change of center position in both horizontal and vertical directions. For the predicted bounding box in the f^{th} frame of trajectory k , we define it as $B_p^{k,f} = (x_p^{k,f}, y_p^{k,f}, w_p^{k,f}, h_p^{k,f})$ which is the center of x axis, y axis, width and height. Similarly, the corresponding ground-truth is $B_g^{k,f} = (x_g^{k,f}, y_g^{k,f}, w_g^{k,f}, h_g^{k,f})$. Then the center position error is defined in Eq.(3), which is average of the standard deviations of center position deviations in all trajectories.

$$\begin{aligned} e_x^{k,f} &= \frac{x_p^{k,f} - x_g^{k,f}}{w_g^{k,f}}, & \sigma_x^k &= \text{std}(\mathbf{e}_x^k), \\ e_y^{k,f} &= \frac{y_p^{k,f} - y_g^{k,f}}{h_g^{k,f}}, & \sigma_y^k &= \text{std}(\mathbf{e}_y^k), \\ E_C &= \frac{1}{N} \sum_{k=1}^N (\sigma_x^k + \sigma_y^k). \end{aligned} \quad (3)$$

It should be noted that the center position error only evaluates the variance of the normalized center deviation instead of its bias. The underlying reason is that the bias has been implicitly considered in the accuracy metric. In other words, larger bias will always result in lower accuracy. By its definition, if the detections in one trajectory consistently biased towards one direction, we will not penalize them.

3.2.3 Scale and Ratio Error

In the same spirit of center position error, we evaluate the stability of scale and aspect ratio of the detections in one

trajectory. We demonstrate the idea in Fig. 2(d). Specifically, we use square root of the area ratio to represent the scale deviation, and define the ratio of two aspect ratios as aspect ratio deviation. The reason we apply square root to area ratio is that we need to keep the magnitude of each type of deviation same. At last, the final result is defined as the average of standard deviations of scale and ratio deviations among all the trajectories. Formally, we have:

$$\begin{aligned} e_s^{k,f} &= \sqrt{\frac{w_p^{k,f} h_p^{k,f}}{w_g^{k,f} h_g^{k,f}}}, & \sigma_s^k &= \text{std}(\mathbf{e}_s^k), \\ e_r^{k,f} &= \left(\frac{w_p^{k,f}}{h_p^{k,f}}\right) / \left(\frac{w_g^{k,f}}{h_g^{k,f}}\right), & \sigma_r^k &= \text{std}(\mathbf{e}_r^k), \\ E_R &= \frac{1}{N} \sum_{k=1}^N (\sigma_s^k + \sigma_r^k). \end{aligned} \quad (4)$$

Same as the center position error, we also focus on the variance instead of the bias of the scale and ratio deviation in this metric. Consequently, if the detections are consistently larger or smaller than ground-truths, they will not be penalized.

4. Validation Setup

In this section, we will introduce our datasets and base detectors used in the experiments.

4.1. Dataset

Our new evaluation metric needs associations for ground-truth. However, in present VID dataset, there is no such information provided. So we adopt MOT dataset in our experiments. In particular, we use two datasets: The first one is the MOT challenge dataset. Since the annotations for testing set are not available, we compile two years' MOT challenge training dataset for experiments. We use MOT 2016¹ training set for training, and MOT 2015² training set for testing. We also exclude the overlapped videos for testing. The other one is KITTI³. We evenly separate them into training set and testing set. Note that we only use car category in the KITTI dataset due to the reason that the number of annotated bounding boxes for pedestrian category is too small to train a stable detector, especially for deep learning based method.

4.2. Basic Models

Currently there is no unified method for VID. Most of them improve upon existing object detectors for still images. Therefore, in our experiment, we choose the follow-

¹<https://motchallenge.net/data/MOT16/>

²https://motchallenge.net/results/2D_MOT_2015/

³http://www.cvlibs.net/datasets/kitti/eval_tracking.php

ing base detectors for validation: Aggregated Channel Feature (ACF) is one representative work for non-deep learning method, while Faster RCNN[35] is the current leading object detector. We apply the VGG16 model [38](VGG-16) pre-trained on ImageNet classification task as our base CNN. For post-processing of these two detectors, we both apply Non-Maximum Suppression (NMS) with threshold 0.5.

5. Validation and Analysis

In this section, we investigate several representative methods for VID. We briefly divide them into two categories: The first type is to improve the aggregation of the output bounding boxes within single frame. In particular, we test the representative method weighted NMS [12]. The second type is to utilize the temporal context across frames. Here, we benchmark two methods, *i.e.* Motion Guided Propagation (MGP) [18] and object tracking [17]. We conduct ablation analyses to investigate how each component affects the final performance.

Some existing works cannot be benchmarked include: 1) Multi-context suppression [18]: This method is proposed to use correlation among different classes to suppress false positives. However, in the two datasets we use, only one class exists. 2) Rescoring [15, 18]: We use the original implementations of the authors, but cannot get similar satisfactory results as reported in their papers on our datasets. We owe the reason to that MOT datasets often contain crowded scene which makes the trajectory generation fail. We will try to tackle this issue in our future work.

All detection accuracy results are shown in Fig. 3, and detection stability results are shown in Fig. 4, 5, respectively.

5.1. Weighted NMS

In a typical pipeline of object detection, usually a detector scores each proposal first, and then an aggregation method is adopted to suppress the redundant bounding boxes. We argue that improper aggregation will significantly lower both accuracy and stability. The original NMS iterates between selecting the unsuppressed bounding box with highest score and suppressing all bounding boxes with an IOU higher than a given threshold with it. However, we miss the valuable treasures of the suppressed bounding boxes. They still provide useful statistics about the object. So in weighted NMS, rather than only keeping the bounding box with highest score, we weighted average it with all the suppressed bounding boxes by their scores. It is first proposed in [12] to improve the mAP of still image detection. In the sequel, we use WNMS for short. There are also other advanced aggregation method such as [27] which uses a learned function for data for aggregation. Due to limited space, we only benchmark the most representative one

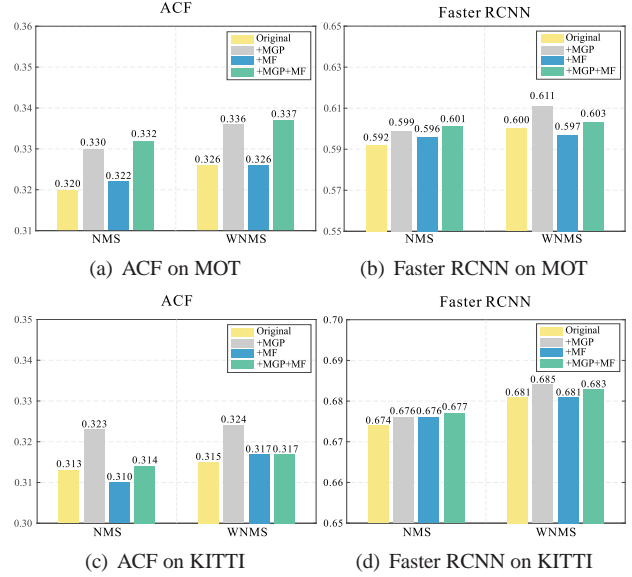


Figure 3. Results of accuracy on MOT and KITTI detected by ACF and Faster RCNN.

weighted NMS.

Result Analysis: Weighed NMS consistently improves over NMS in both accuracy and stability. Notably, the marginal benefit even increases: The gap of these two methods for Faster RCNN is even larger than that of ACF. We owe the reason to the increased complexity of the detection system: A complicated system may reduce the bias, however it may not reduce, even increase the variance of the outputs. Weighted NMS effectively makes up this disadvantage by averaging over samples. Besides its effectiveness, it is also easy to implement and almost cost free.

5.2. Motion Guided Propagation

Due to the high correlation of adjacent frames, propagating detections to adjacent frames may help recover false negatives. Motion Guided Propagation (MGP) is proposed to alleviate such issue in [18]. MGP takes the raw bounding boxes before aggregation, and then propagates them bidirectionally using optical flow. The propagated bounding boxes are treated equally as other detections, and are used in the subsequent aggregation. Different from the original paper, we empirically find that adding a decay factor for each propagation could improve the results. The decay factor is dataset dependent, and we tune it using validation set.

Result Analysis: MGP consistently improves the accuracy for both ACF and Faster RCNN. In stability, MGP is especially helpful for fragment error. This is reasonable since MGP propagates detections with high confidence bidirectionally. It is not surprising that it helps to recover the

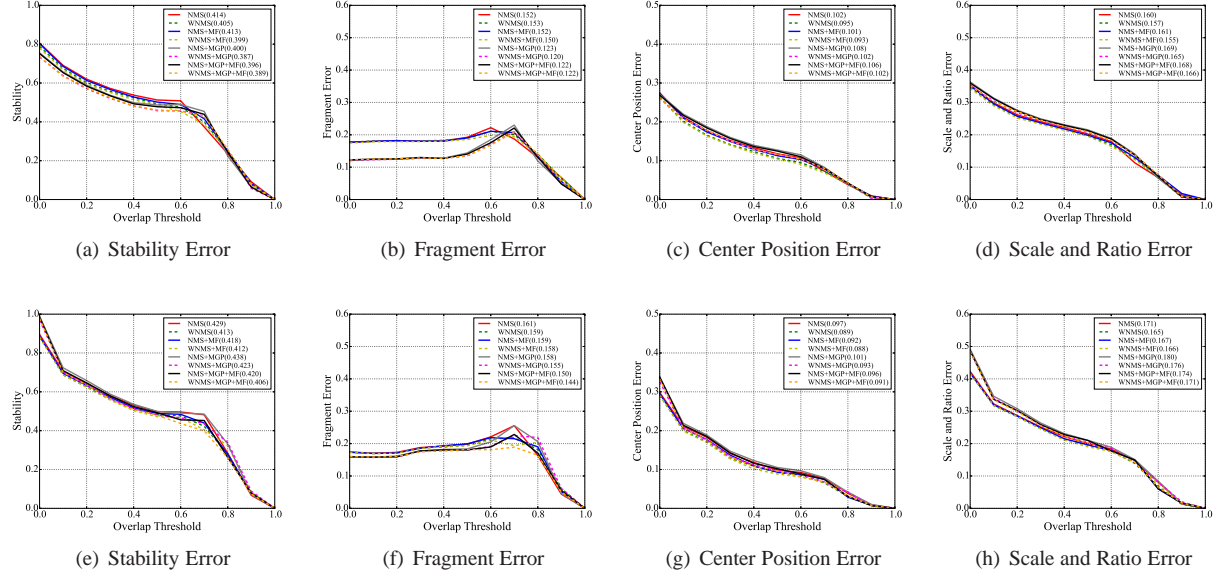


Figure 4. Results of stability error of ACF. (a)-(d) on MOT, (e)-(h) on KITTI

false negatives. As for center error and scale and ratio error, the impact is uncertain on different datasets.

5.3. Object Tracking

We next investigate the use of object tracking method to smooth the trajectory. In our implementation, we choose Median Flow (MF) [17]. It is proven to be an efficient and effective short term tracking method. The merit of it lies in that it can detect self failure reliably by checking forward-backward error. Different from MGP, we apply MF after the bounding box aggregation phase (NMS or weighted NMS). Moreover, we only use it to smooth the detection bounding box, and do not alter the detection score or add new detection bounding boxes. Concretely, we start tracking with high confident detections (detection score higher than 0.8 in our experiments). If the tracker reports a reliable tracking result, we find the bounding box with highest IOU (should be at least higher than a given threshold, 0.5 in our case.) with it, and then average them as the final bounding box. For the detections without any associated tracking bounding box, we keep it unchanged.

Result Analysis: As discovered in the evaluation, median flow is effective at stabilizing the detections, but has minor effect on accuracy. Even though median flow is one simple tracker and we only average the results, tracking is still beneficial to the performance. We believe that a more sophisticated tracker and better fusion method will further improve the results.

5.4. Methods Combination

Finally, we put together all these improvements. The best combination consistently improves over the baseline 1.0~2.0 in accuracy and 0.03~0.06 (8%~15% relative improvement) in stability. We further illustrate some visual results in Fig. 7. It is easy to observe that the combined method localizes the objects of interest more accurately and consistently compared to the baseline, especially in the crowded and occluded scene. For more intuitive comparisons, we refer the readers to the videos in supplemental material.

6. Metric Analysis

In this section, we investigate the relationship between accuracy and stability in the proposed metric. Through the analysis, we justify that these two metrics are complementary and both necessary to assess the performance of a detector in VID.

6.1. Correlation Analysis

We first analyze the correlation between all the metrics including the accuracy and three types of stability. We run Faster RCNN and ACF with 8 different methods on both 7 video sequences of MOT and 9 video sequences of KITTI. As a result, we have 256 samples for each metric. We plot the absolute value of the correlation matrix in Fig. 6. For each cell in the figure, darker color denotes higher correlation between corresponding metrics.

It is easy to see that the accuracy metric has relatively low correlation with other three stability metrics. This is anticipated because they should characterize different aspects

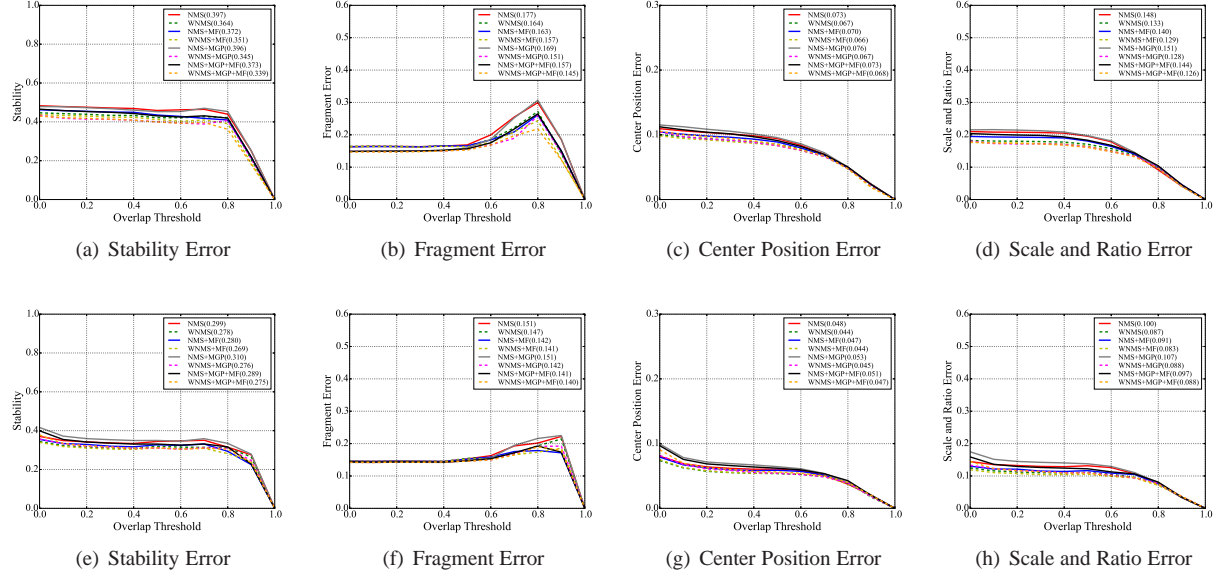


Figure 5. Results of stability error of Faster RCNN. (a)-(d) on MOT, (e)-(h) on KITTI

of quality in VID. Thus it is meaningful to measure both of them. Then we take a closer look at the three stability metrics. Among them, center position error and scale and ratio error are the most correlated, and fragment error is less correlated with them. This is reasonable since center position error and scale and ratio error both represent the spatial stability in one trajectory, while fragment error represents the temporal stability.

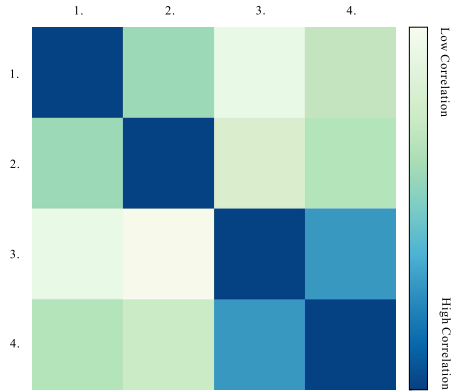


Figure 6. Correlation matrix for four measurements. 1: detection accuracy, 2: fragment error, 3: center position error, 4: scale and ratio error. Best viewed in color.

6.2. Accuracy vs Stability

In order to interpret the trade-off between accuracy and stability, we draw the scatter plot of accuracy and stability of various methods in Fig. 8. The values are mAUC of accuracy curve and stability curve, respectively. The trends on these two datasets of two detectors are similar. Interest-

ingly, we find that there is no single best method that outperforms others in both accuracy and stability. Specifically, weighted NMS boosts both these two metrics, while MF mostly improves the stability and MGP improves the accuracy. Although MF and MPG both utilize motion information to guide detection, the impact on performance differs and complements. In practice, how to compromise between these two aspects relies on the application at hand.

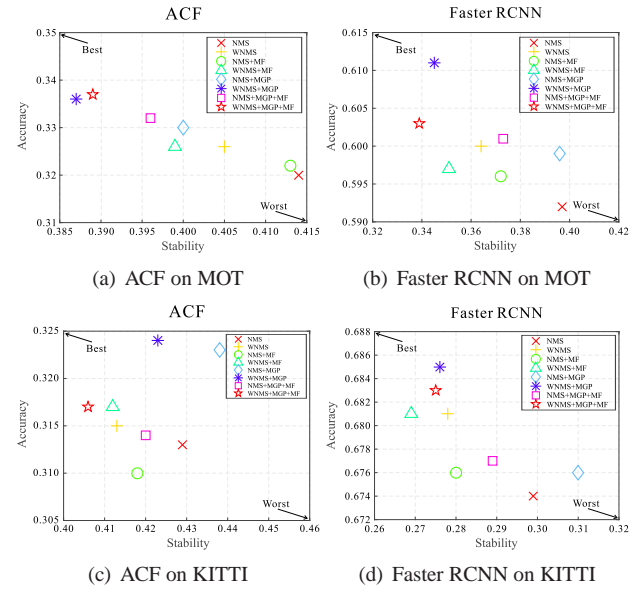


Figure 8. An accuracy-stability visualization for all methods on MOT and KITTI detected by ACF and Faster RCNN



Figure 7. Selected examples on MOT and KITTI using NMS and our best combination of methods. The base detector is Faster RCNN.

7. Discussion and Future Works

One limitation of our work is that we need associations to evaluate the stability. Our method essentially finds the best possible associations using ground-truths. Consequently, we cannot use standard VID dataset such as ImageNet VID [8] or Youtube dataset [33] in our experiments. However, it usually takes a long time to annotate each trajectory in the videos. We will try to investigate how to reduce such labor works in our future work.

Currently, most VID methods rely on the explicit associations to refine the final results. This kind of methods actually blurs the boundary of VID and MOT: They can also output the associations between detections easily. However, this is not the only way to consider temporal context. Very recently, two concurrent works [43, 10] tried to use the Conv-LSTM framework to address the temporal continuity in video segmentation. They could enjoy all benefits of end-to-end learning without explicitly calculating motion or associations between frames. Nevertheless, seeking such a method for VID is still an open problem. Last but not least, though we have proposed a metric to evaluate the stability in VID, we still cannot model it directly into learning. How to integrate the stability error into VID and MOT formulation is also an interesting direction to pursue.

8. Conclusion

In this paper, we have investigated one important missing component in current VID and MOT evaluation – stability. First, we analyzed the sources of the instability, and further decomposed it into three terms: fragment error, center

position error, scale and ratio error. For each term, we proposed its corresponding evaluation metric. These proposed metrics are intuitive and easy to measure. Next, we conducted comprehensive experiments to evaluate several existing methods that improve over still image detectors based on the new metrics. Through empirical analyses, we justified that accuracy and stability are complementary. Both metrics are necessary to characterize the performance of a detector in VID. Furthermore, we have demonstrated the trade-off between accuracy and stability for existing VID methods. We wish our work could inspire more subsequent works that address the stability issue in VID.

References

- [1] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.
- [2] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016.
- [3] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1806–1819, 2011.
- [4] K. Bernardin and R. Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):1–10, 2008.
- [5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

- [6] W. Choi, C. Pantofaru, and S. Savarese. A general framework for tracking multiple people from a moving camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1577–1591, 2013.
- [7] E. Dagan, O. Mano, G. P. Stein, and A. Shashua. Forward collision warning with a single camera. In *Intelligent Vehicles Symposium*, 2004.
- [8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [9] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *CVPR*, 2014.
- [10] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, R. Klette, and F. Huang. STFCN: Spatio-temporal fcnn for semantic video segmentation. *arXiv preprint arXiv:1608.05971*, 2016.
- [11] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained Part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
- [12] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware CNN model. In *ICCV*, 2015.
- [13] R. Girshick. Fast R-CNN. In *ICCV*, 2015.
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
- [15] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang. Seq-NMS for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.
- [16] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.
- [17] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *ICPR*, 2010.
- [18] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, et al. T-CNN: tubelets with convolutional neural networks for object detection from videos. *arXiv preprint arXiv:1604.02532*, 2016.
- [19] K. Kang, W. Ouyang, H. Li, and X. Wang. Object detection from video tubelets with convolutional neural networks. In *CVPR*, 2016.
- [20] Z. Khan, T. Balch, and F. Dellaert. MCMC-Based particle filtering for tracking a variable number of interacting targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1805–1819, 2005.
- [21] M. Kölsch and M. Turk. Robust hand detection. In *FGR*, 2004.
- [22] T. Kong, A. Yao, Y. Chen, and F. Sun. HyperNet: Towards accurate region proposal generation and joint object detection. In *CVPR*, 2016.
- [23] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
- [24] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In *CVPR*, 2016.
- [25] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005.
- [26] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, 2009.
- [27] S. Liu, C. Lu, and J. Jia. Box aggregation for proposal cecimation: Last mile of object detection. In *ICCV*, 2015.
- [28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In *ECCV*, 2016.
- [29] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72, 2014.
- [30] E. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *FGR*, 2004.
- [31] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *CVPR*, 1997.
- [32] E. Osuna, R. Freund, and F. Girosit. Training support vector machines: An application to face detection. In *CVPR*, 1997.
- [33] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.
- [34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. 2016.
- [35] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*.
- [36] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.
- [37] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.
- [38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014.
- [39] G. P. Stein, O. Mano, and A. Shashua. Vision-based ACC with a single camera: Bounds on range and range rate accuracy. In *Intelligent Vehicles Symposium*, 2003.
- [40] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *NIPS*, 2013.
- [41] S. Tripathi, Z. C. Lipton, S. Belongie, and T. Nguyen. Context matters: Refining object detection in video with recurrent neural networks. In *BMVC*, 2016.
- [42] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal on Computer Vision*, 104(2):154–171, 2013.
- [43] S. Valipour, M. Siam, M. Jagersand, and N. Ray. Recurrent fully convolutional networks for video segmentation. *arXiv preprint arXiv:1606.00487*, 2016.
- [44] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal on Computer Vision*, 57(2):137–154, 2004.
- [45] B. Yang, J. Yan, Z. Lei, and S. Z. Li. CRAFT objects from images. In *CVPR*, 2016.
- [46] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.